【原著】

作文試験におけるコンピュータの利用について

—全米学力調査 NAEP での実施例を中心として—

石岡 恒憲（大学入試センター）

従来の紙筆テストに替わって行なわれつつあるコンピュータによる作文テストについて，全米学力調査(NAEP) を例にとり，その仕様や電子化の背景，及び自動採点されない理由について論考する。また NAEP で実施されていない作文の自動採点について，以前 NAEP 作文テストの採点に使われていた E-rater と，今後使われることが予想される IntelliMetric の仕様を紹介する。NAEP における次回の作文テストは 2017 年に行われることが既に決定しているが，IntelliMetric の ESL e-Write の設問とのすり合わせが自動採点に向けた今後の課題になっている。

1 はじめに

2011 年の全米学力調査 (National Assessment of Educational Progress, NAEP)では，8 年生と 12 年生に対して，作文テストが従来の紙筆テストに換わり，初めてコンピュータによって実施された。そのテストは，単に従来の紙と鉛筆をコンピュータに置き換えただけのものではない。現在のデジタル技術が十分に活用できるよう，ワープロ機能や短いビデオや音声の活用を視野に入れているとしている。事実，実際に公開されている 8 年生に対する試験問題や実施手順を見るに，取り扱う作文のタイプには，テキストによる問いかけのほか，写真を含むもの，音声によるもの，ビデオを見て問いかけに答えるものの 4 つがある。加えて，多くの学生が受験できるようユニバーサルデザインについての配慮がなされている。たとえば，問題文についての音声読み上げやフォントサイズの変更が可能である。また電子上のスペルチェックが利用できる。2012 年には，4 年生に対してもコンピュータを用いた作文の試行テストが予定されている。

ここで NAEP について少し解説をしておく。NAEP は約 2,000 の学校から約 10 万人の学生を対象に行うサンプリング調査である。NAEP には大別して Main NAEP と Long-term Trend NAEP がある。Main NAEP は社会や時代の変化に応じた教育課題に焦点を当てた調査で，2011 年の作文はこの Main NAEP で実施した教科の一つである。この年は他に，数学，読解，理科についての調査が実施された。他の年ではたとえば，公民，地理，合衆国史，芸術といった教科が選ばれている。教科は毎年変わる。一方，Long-term Trend NAEP は，スペルや四則演算のように時代の変化に関わらず不変に求められる基礎学力を調査する。長年にわたって同じテスト項目を使い，継続的な傾向から学力の変化を探ることを目的としている。隔年に実施されている。他に State NAEP というものがあり，希望する州が Main NAEP にリンクさせながら同じ問題を使って，各州の学力状況を詳細に測る目的で実施する。

さて，NAEP はいままで理科の実験など一部を除き，基本的には紙と鉛筆による形式で実施されてきた。しかし，実施団体である National Assessment Governing Board, NAGB はコンピュータによる評価方法を指向しており，2011 年度の作文テストが，初めての完全（fully）なコンピュータによる試験になっている。完全でないものとしては，同

年 2011 年の数学で能力に応じて問題が出題される適合型テストが，一部の学生に対して実施された。また理科においては 2009 年に長時間を要する実験の観察を，コンピュータによる疑似シミュレーションとして実施した。コンピュータ利用化が進んだ背景としては，小学生でさえもキーボード操作によるコンピュータ利用に習熟していることが挙げられる。NAEP(2012)の調査によれば，2009 年には 4 年生の 89%が自宅にコンピュータ (home computer)を持っており，読解や学校での語学の宿題にコンピュータを使っているとしている。

本稿では，この初めてのコンピュータ化された作文試験について，その仕様を示すとともに，今後の方向性や果たすべき課題について論考する。2 節では，NAEP 作文試験の，受験者からみた仕様について，スクリーン・ショットを随時示しながら紹介する。3 節では，NAEP 作文試験の評価の観点や採点方法について説明する。また，特記すべきことであるが，NAEP の作文試験は 2000 年以降，人ではなくコンピュータ(E-rater システム)によって自動採点されていたにもかかわらず，試験自体がコンピュータ化された 2011 年では，自動採点が用いられていない。その理由や背景について 4 節に述べる。5 節では，まとめと今後について述べる。

2　NAEP 作文試験の見た目

2.1　2部構成

NAEPの作文試験は2部構成であり，第 1 部は，作文 2 題をそれぞれ30分で解答する。問題はすべてコンピュータ上で出題される。

設問には，テキストによる問いかけのほか，写真を含むもの，音声によるもの，ビデオを見て問いかけに答えるものの 4 つがあるが，釣り合い型不完備ブロック計画(Balanced incomplete blocks design; BIB design)に基づいて，このうち2つが出題される。不完備というのは4つから2つを選ぶ組み合わせの一部のみが指定されることによる。被験者によって異なる組み合わせのブロックを与えることで，NAEPテストに割く時間を最小限に抑えながら，広範囲な学力を正確に測ることを目的としている。

第2部はアンケートであり，解答者の属性，家族，教育についての28の設問に解答する。例えば，解答者の人種（白人/黒人/アジア人/アメリカン・インディアンorアラスカ・ネィティブ/ネィティブ・ハワイアンorポリネシア）や新聞・雑誌の購読，蔵書数，自宅でのコンピュータ利用の有無などについて，回答する。コンピュータでの回答に際して，必要な操作は全て回答者が行う

2.2　解答における操作

図 1 は写真を含むテキストをみて，解答する問題であるが，解答ウィンドウの左には 5 つのアイコンが示される。図 1 ではこのアイコンを，画面左端に拡大して表示している。上から，「ボリューム調整」「読み上げ」「フォントサイズ調整」「マーカー」であり，左下にテストの残り時間を表示する「時計」の各ボタンが用意されている。文字の部分を選択(クリック)し，「読み上げ」ボタンを押すことで，選択部分を読み上げてくれる。音量は「ボリューム調整」ボタンのスライダーを用いて調整する。

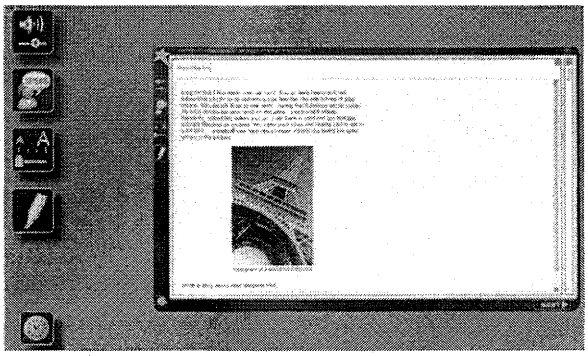

図 1 ：設問表示で使える各種ボタン

図 2 は，「フォントサイズ調整」のスライダーを用いて，フォント(付随して写真も)を拡大しているところである。

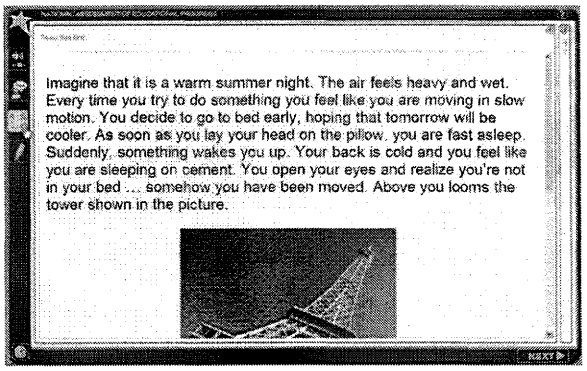

図２：フォントサイズを最大にする

「マーカー」ボタンは，文字の部分を選択しこのボタンを押下することで字色を黄色に変えることができる。「時計」ボタンはこれをクリックすることで，試験の残り時間が表示される。

図３は写真を含むビデオをみて，解答する問題であるが，ビデオ画面中の矢印ボタンを押下することで試験は始まる。ビデオ音声についてはクローズド・キャプションにより文字で表示することできる。ビデオ画面の下にある設問文については，テキスト問題と同様に，「読み上げ」や「マーカー」等が利用できる。

図３の画面右側では，解答するワープロ画面を表示させ，ビデオを見ながら，解答することができる。ビデオは，適宜，止めたり，意図する箇所に移動したりすることができる。

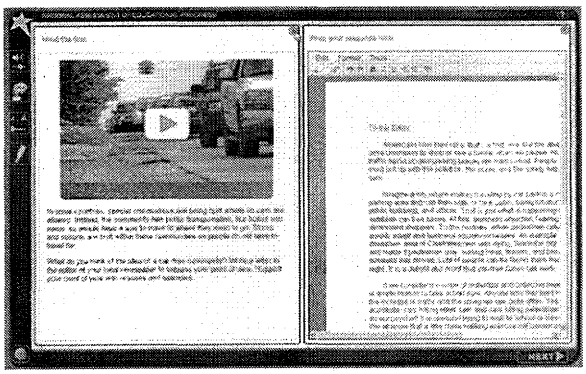

図３：ビデオを見て解答する

ビデオ画面を閉じて，ワープロ画面のみを全画面に切り替えることもできる（図４）。

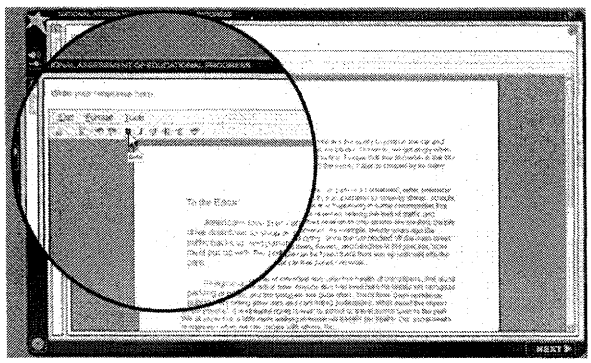

図４：Wordに似たワープロ画面

このワープロ画面は，Microsoft Wordに見た目が似ており，カット&ペーストや，「繰り返し」「元に戻す」ができる。また，太字やイタリックなどの簡単な文字修飾，インデント/アウトデントに加えて，スペルチェックのボタンが用意されている（図４）。このボタンによる操作は全て，プルダウンメニューによる代替操作が可能である。

2.3 ユニバーサルデザイン

この作文解答において特筆すべきは，障害者に対する配慮である。前項でみてきたように，出題文の音声読み上げや，「フォントサイズ調整」ボタンで 48 ポイントサイズにまで拡大したり，「マーカー」ボタンで地色を黄色にしたりすることができる。NAEP では障害者に対しても代替問題を使用せず，健常者と同じ問題を解答させることを基本としている。そのやり方が公平性を担保し，障害者の尊厳を損ねず，また作題や採点の手間を大幅に軽減するとしている。

3 NAEP 作文テストの評価

3.1 出題内容

受験者は 2 題を各 30 分で解答するが，作文テストでは，以下に示す３つのコミュニケーションの目的における作文を評価する。

- [説得]すること(To persuade)
- [説明]すること(To explain)
- [経験や体験]を述べること(To convey

experience)

作文テストの枠組について述べてある Writing Framework(NAEP, 2011)によれば，これら３つの比率は表１の通りである。これより，学年が上がるにつれて，説得することと，説明することの比率が増えてきていることがわかる。また出題は，この比率に応じて被験者に提示されていると予想される。

表１：コミュニケーション能力測定の割合（学年別）

目的	４年生	８年生	12 年生
説得	30%	35%	40%
説明	35%	35%	40%
経験や体験	35%	30%	20%

3.2 採点基準

NAEP の責任主体は前述のとおり NAGB であるが，作文テストにおける枠組み(Driscoll, 2011)については，契約により全米の有力なテスト機関である ACT Inc.に一任されている。NAEP では採点スコアは多くの事例に基づいて予め使われた採点済の訓練データを通して採点される。作文の特性に対する評価観点は以下の３つである。これらは 3.1 節で述べた３つのコミュニケーションの目的に関わらない。

- アイディアの展開：Development of Ideas
- アイディアの組織化:Organization of Ideas
- 適切な語彙，字句の使用：Language Facility and Use of Conventions

スコアは基準表を用い，上記の３つの特性に応じた評価項目に基づいて，その総合点が１点から６点で報告される。６点が満点で１点が最低点である。ただし，あまりに分量の少ないもの，英語で書かれていないもの，問いかけの内容と話題が大きく外れているものについては０点が与えられる。この作文の採点は，訓練を受けた複数の人間の採点に基づいて評定される。

4 コンピュータによる自動採点

4.1 代表的なシステム

2000－2007 年，NAEP の作文試験の採点に使われていた E-rater と 2011 年には使われなかったが，今後使用されることが予定される IntelliMetric について説明する。

(1) E-rater

E-rater は世界最大のテスト機関である Educational Testing Service, ETS の Burstein らの研究グループによって開発されたシステムであり，2004 年に新バージョン(Ver.2.0) が開発された(Attali & Burstein, 2005)。E-rater では複数の言語上の特徴量に基づく重回帰によってスコアを計算するが，Ver.2 では用いられる変量の数が Ver.1 時代の 60 余りからわずか 12 に厳選され，論題に依らずに固定となった。その変量の示す特徴量自体も，良い作文(good writing)を示す性質と関係がより付くように改良されたとしている。これは従来,tricked と批判されていた採点の仕組みを簡素化し，採点結果への説明性(accountability)を明確にするためである。

その 12 の変量は，1.総ワード数に対する文法エラーの割合，2.総ワード数に対する語の使用法についてのエラーの割合，3.総ワード数に対する手順のエラーの割合，4.総ワード数に対するスタイルについてのエラーの割合，5.談話(discourse)ユニットの数，6.各ユニットにおける平均ワード数，7.当該エッセイの 6 点法によるコサイン類似度（評価対象である作文と比較作文との２つのベクトルのなす角のコサインの大きさ;統計学での相関係数）が最大となるスコア点，8.最高点(通常 6 点)を得たエッセイとのコサイン類似度，9.単語の繰り返しの程度を示す指標：総単語数(token)に対する異なったワード種類(word type)の割合，10. Breland (Breland et al, 1994)のワード頻度指標に基づく語彙の困難度，11.平均単語長さ，12.単語の総数，である。

これら 12 変量に係る重み付けは経験則に

よって定められる。TOEFL エッセイについては,それぞれの重みは順に 0.05, 0.02, 0.07, 0.08, 0.21, 0.12, 0.04, 0.07, 0.08, 0.03, 0.03, 0.20 であるとしている(Attali & Burstein, 2005)。

(2) IntelliMetric

IntelliMetric は Vantage Learning 社によって,エッセイや自己完結型(open-ended)問題の採点のために開発されたシステムであり,開発者サイドが自称するところの知能に基づいたモデルに基づいて情報処理理解を行なっている。技術的な背後にあるのは,人工知能,ニューラルネット,計算機言語学であるとしている。与えられた論題に対して,IntelliMetricは生徒の回答から400もの特徴量を抽出し,スコア推定に有効な特徴量を抽出し,スコアモデルに係る重みを推定する(Elliot, 2003)。

IntelliMetric による評価スコアの観点は,文献によって多少の違いがあり,また用いられているワーディングも一貫していないが,概ね以下の 5 つである:1.目的や主題に対しての結束性や一貫性,2.内容の幅や発想の展開,3.論旨の展開や文章構成,4.文の完全性や多様性,5.英語のルールへの適合。

上記 5 つの評価スコア観点と(前述の)約 400 の特徴量への関係づけの対応は,ある 1 つの特徴量が排他的に 1 つの評価スコアに対応するのではなく,複数の評価スコアに対応しうるし,また逆に 1 つの評価スコアに対応する特徴量は複数存在する,いわゆる多対多の関係である(Elliot, 2003)。

4.2 IntelliMetricが自動採点に使われなかった理由

2007 年 3 月 14 日の Education Week 誌の記事にもあったように,2007 年の時点で既に NAEP の新しい試験の枠組みが ACT Inc.によって検討されていた。日程的にも 2011 年の実施には余裕があると思われたことから,NAEP でもエッセイの採点は GMAT 同様に IntelliMetric によって採点されるものだと著者は予想していた。その実現を妨げる技術的な要因がもしあるとすれば,それは,ACT Inc.が高校生に提供する(そして人間が採点する)COMPASS e-Write(英語を母語とする高校生レベルの作文問題)及び ESL e-Write(英語を第 2 外国語とする高校生レベルの作文問題)の評価観点が,いずれも話題の焦点(Focus),内容(Content),組織化(Organization),文体(Style),慣例(Conventions)の 5 つで,NAEP の 3 つとは一致しないことであった。また COMPASS が評価項目のそれぞれの項目に対して 35%,10%,15%,35%,5%の重み付け採点,いわゆる積み上げをするのに対し,NAEP の採点は全体についての印象で評価がされる総合評価であることもその適用を難しくする要因とは思われた。

ACT(2008)には「どのようにして COMPASS 及び ESL e-Write での採点スコアを IntelliMetric に学習させるのか?」との問いとそれに対する回答がある。これによれば ACT Inc.は Vantage Learning 社に対して COMPASS e-Write に対して 1 課題あたり 300 件の,また ESL e-Write に対して 1 課題あたり 500 件の回答文とその評価結果を渡し,スコアのすり合わせを依頼しているという。もし IntelliMetric の採点が,訓練を受けた人間の複数の採点のバラツキの範囲内でなければ,さらなる調整が必要であるし,IntelliMetric の信頼性や妥当性が要求したレベルに達しないならば,ACT Inc.はもとの人間の採点結果や課題文そのものを精査し,コンピュータと人間との採点の違いがどこにあるのかを探求しなければならないとしている。そしてその違いが修正できないなら,COMPASS あるいは ESL e-Write の課題文を IntelliMetric で採点することはできないとしている。ACT Inc.の提供する ESL e-Write の論題が,改変されずにそのまま NAEP の 12 年生の作文に用いられていることを考慮

すれば，結局，2011 年の時点では IntelliMetric の採点が ACT Inc.の所持する課題文に対して必要な水準を満たすことができなかったと予想される。

5 おわりに

NAEP(2011)の 2017 年までのスケジュール表によれば，作文テストに関しては 2011 年（8,12 年生）以降，最も近いのは 2017 年（4,8,12 年生）である。作文は数学や読解などと並んで最も頻繁に採用される教科で，過去には，4－5 年おきに実施されてきたのであるが，次回まで 6 年の余裕があり，次回はコンピュータによる自動採点が行われるのは確実であると思われる。IntelliMetric は，アメリカの医学大学院進学のための適性試験 MCAT (Medical College Admission Test)の作文試験の採点にも 2007 年より用いられており，さらにはカレッジ入学のためのインターネットベースのテストである ACCUPLACER プログラム中の作文テストである WritePlacer Plus でも使われているからである。

我が国でも，著者らの開発した Jess など，作文の自動採点は技術的には可能である。またその妥当性についても定量的な検証がされている。しかしながら日本語の作文においては，キーボード操作に「かな漢字変換」が混じるために，自宅や学校での操作環境を同一に保つことが難しいと予想される。また欧米とは違い同時一斉に試験を実施する要請から，そのインフラの整備にも困難が伴う。このため，コンピュータを用いたテキストの入力や書式付きデータの入力，およびそれらを入力データとする自動採点については，現時点ではハードルが高いと思われる。

もし我が国でコンピュータによる作文テストが実施されるとしたら，それは手書きの作文を画像として読み取り，手書き文字認識をすることなく，画像そのものを電子データ化し，採点者へ配布，採点結果を返却，データベースに格納する形で進むと予想される。採点に際しては，採点者の確保や採点基準の策定など新たな問題が生じるが，技術的にはデータ転送時に第 3 者にデータを盗み見られることのないように，またデータの改ざんがされないように，十分なセキュリティを確保することが重要である。

テストにおけるコンピュータの利用は，時代の趨勢であり，作文のみが例外にはなり得ない。その仕組みの構築は喫緊の課題であり，もはやその検討の時期にきていることは間違いないであろう。

参考文献

ACT (2008). Answers to Frequently Asked Questions about COMPASS e-Write & ESL e-Write, ACT, Inc. http://www.act.org/compass/pdf/ewritefaq.pdf

Attali, Y. & Burstein, J. (2005). Automated essay scoring with e-rater v.2.0 (ETS RR-04-45), Princeton, NJ: Educational Testing Service.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). The College Board vocabulary study (College Board Rep. No. 94-4; ETS RR-94-26). New York: College Entrance Examination Board.

Driscoll, D.P, Avallone, A.P., Orr, C.S., and Crovo, M. (2011). Writing Framework for the 2011 National Assessment of Educational Progress, National Assessment Governing Board.

Elliot, S. (2003). IntelliMetric: From Here to Validity, 71-86. In Shermis, M. & Burstein, J. eds. Automated essay scoring: A cross-disciplinary perspective. Hillsdale, NJ: Lawrence Erlbaum Associates.

National Assessment of Educational Progress NAEP (2011). http://nces.ed.gov/nationsreportcard/

National Assessment of Educational Progress NAEP (2012). NAEP Writing Computer-Based Assessment, An Overview for Grade 4.

The 2009–2010 National Assessment Governing Board (2011). Writing Framework 2011. http://www.nagb.org/publications/frameworks/writing-2011.pdf